非典型 DNA 位点 G4 探测方法综述

Review：Detection methods for Non-canonical DNA structures G4

生物信息 张佳胜 指导老师 张勇 刘晓兰

摘要：

四链 DNA 结构，又被称为 G4，通常由鸟嘌呤富集的序列通过特殊的拓扑结构形成。科学家发现，这种结构广泛存在于体内基因组上，并且在一些关键性的区域承担着极其重要的功能。为了探究 G4 的位点相关性，研究人员需要鉴定出基因组上存在的 G4 结构以及找出可能的 G4 位点。随着研究的进展，逐步发展出了以二代测序技术为核心的 G4 ChIP-seq 以及 G4-seq 技术，能够高通量地识别基因组层面的 G4 位点，同时，计算生物学家采用多种算法，针对基因组序列进行预测分析。本综述的核心是 G4 探测方法，将从以下几个方面展开：首先介绍 G4 结构的特点，然后我们逐步分析 G4 的测序技术以及计算预测方法，最后我们将总结 G4 的研究进展。

关键词：G4，G4 ChIP-seq，G4-seq，正则表达匹配，滑动窗口和打分，机器学习

ABSTRACT

Four-stranded DNA structures called G-quadruplex (G4) arise from guanine-rich sequence through special topological structures. Scientists find that G4s located widely in the genome in vivo, and especially perform important roles in critical regions. To explore this correlation between G4 with gene regions, researchers need detection and identify the potential G4s from genome. With the development of studies, G4 ChIP-seq and G4-seq, whose core is the next-generation sequencing method, are gradually maturing, which means that we can sufficiently high-throughput to scan and identify G4s on a genomic level. Meanwhile, in silico, computational biologists predict the G4 locations and analysis these sequence from genome assembly by using many computational methods. The present review aims at G4 detection methods and we will discuss it from following aspects: first, start with the feature and structure of G4. Next, we introduce the G4 sequencing technologies and computational prediction methods in turn. Finally, we will conclude the progress of G4s.

Key words：G4，G4 ChIP-seq，G4-seq，regular expression matching algorithm，sliding windows and scoring，machine learning

1 引言

通常，基因组上的绝大部分 DNA 表现为右手双螺旋结构，我们称之为典型性 DNA (B-form

DNA, B-DNA)。但是事实上，DNA 的二级结构多种多样，存在一部分特殊的 DNA 结构，研究者们称之为非典型 DNA[1] (non-B-form DNA)。非典型 DNA 包括 Z-DNA、十字形 DNA (cruciforms)、三链体 DNA (triplexes)以及四链体 DNA (G4)等。尽管在很长一段时间内，这些结构都被认为只存在于体外的 DNA 序列上，但随着研究的进展，越来越多的证据表明，非典型 DNA 存在于体内并且承担着重要的生物学功能。由于非典型 DNA 的特定作用，使得它也作为基因组结构研究的一个重要方面，被研究者们广泛关注。

G4 (G-quadruplexes)，又被称为鸟嘌呤四链体，是非典型 DNA 结构之一。这种结构与部分关键性的基因区域有关，尤其是富集在启动子、端粒等区域，这可能意味着 G4 在体内承担着潜在的重要功能。多项研究结果表明[2][3]，G4s 与 DNA 的复制，转录、基因表达、端粒的稳定性以及表观遗传记忆等都具有一定的相关性。本文将从 G4 的结构特点出发，逐步介绍 G4 的探测技术研究进展以及计算预测手段，最后我们将进行展望和总结。

2 G4 及其基本特征

G4 是能够折叠成四链 DNA 结构的单链鸟嘌呤富集 DNA 序列[4]，一般产生于 G 四平方面的自我堆叠，这种 G 四平方面通常由四个鸟嘌呤通过环 Hoogsten 氢键连接，排列而成。在这个 G 四平方面的中心，存在一个一价阳离子（通常是 K+或者 Na+）稳定结构，而在这些四平方面之间，通常由受挤压的单链 loop 序列连接。G4 存在多种空间拓扑结构，取决于 G4s 堆叠 G 四平方面的数量、中间相互连接 loop 的长度与 loop 的序列、折叠时核酸链的方向以及中心通道阳离子的性质（见图 1）。由于 G4 的空间拓扑结构特殊，它的热稳定性通常比较高，不利于 DNA 链的开合以及与其他分子的交流。这使得 G4 存在于特定的基因区域会阻碍或者关闭特定的功能活性。

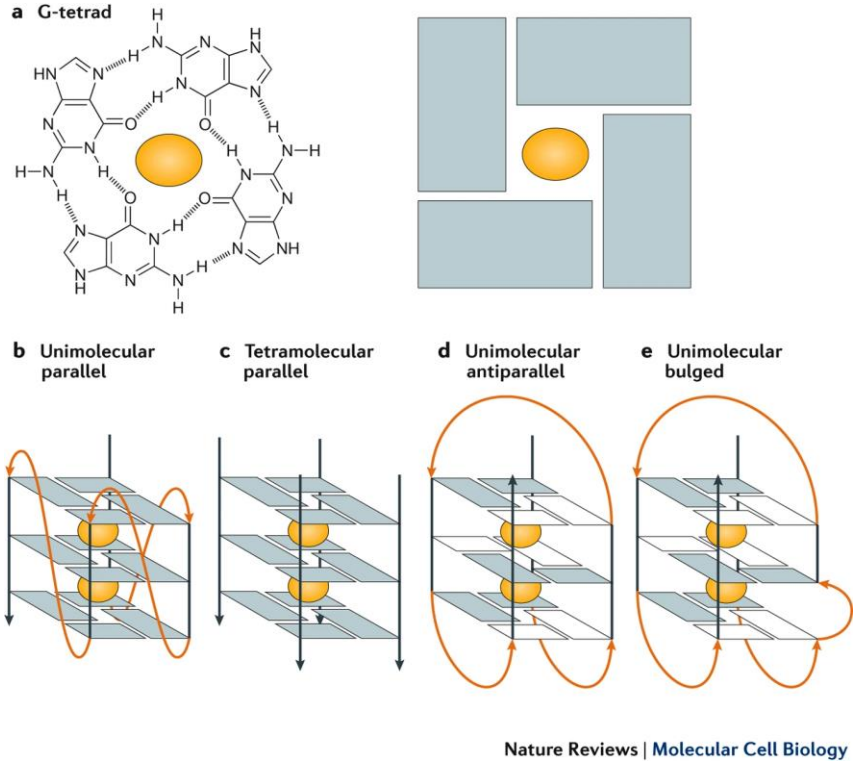

Figure 1 | G4s 结构[4]。

由四个鸟嘌呤通过环 Hoogsten 氢键排列而形成四平方面，四平方面的中心由金属阳离子稳定结构，然后通过链的不同排列方式将四平方面堆叠，构成所谓的 G4s。图 a-e 黄色椭圆为金属阳离子，图 b-e 矩形为鸟嘌呤，黑色线条表示鸟嘌呤核苷酸 runs，黄色线条表示中间连接的 loop。

研究发现[1]，G4 结构存在于多种生物体内，其中人类基因组上已发现有超过 375, 000 条 G4s 序列。G4 在基因组上的位点并不随机，而且通常在近缘物种之间高度保守。有趣的是，大部分的 G4 具备相同的序列模式，我们称之为典型 G4s (canonical G4 sequence)，典型的 G4s 基序如右所示：$G_{\geqslant 3}N_{1\text{-}7}G_{\geqslant 3}N_{1\text{-}7}G_{\geqslant 3}N_{1\text{-}7}G_{\geqslant 3}$。当然，也存在部分非典型的 G4 序列，它们的 loop 长度可能大于 7 或者鸟嘌呤 runs 数目大于 4。

3 G4 的探测手段

为了更深入地了解 G4 在体内的作用，探测 G4s 在人类基因组上的分布和位点成为了一个关键性的问题。目前已经掌握并在使用的 G4 探测手段主要分为三类——高通量测序(G4 ChIP-seq 以及 G4-seq)，传统的物理手段以及计算机预测的方法[5]。

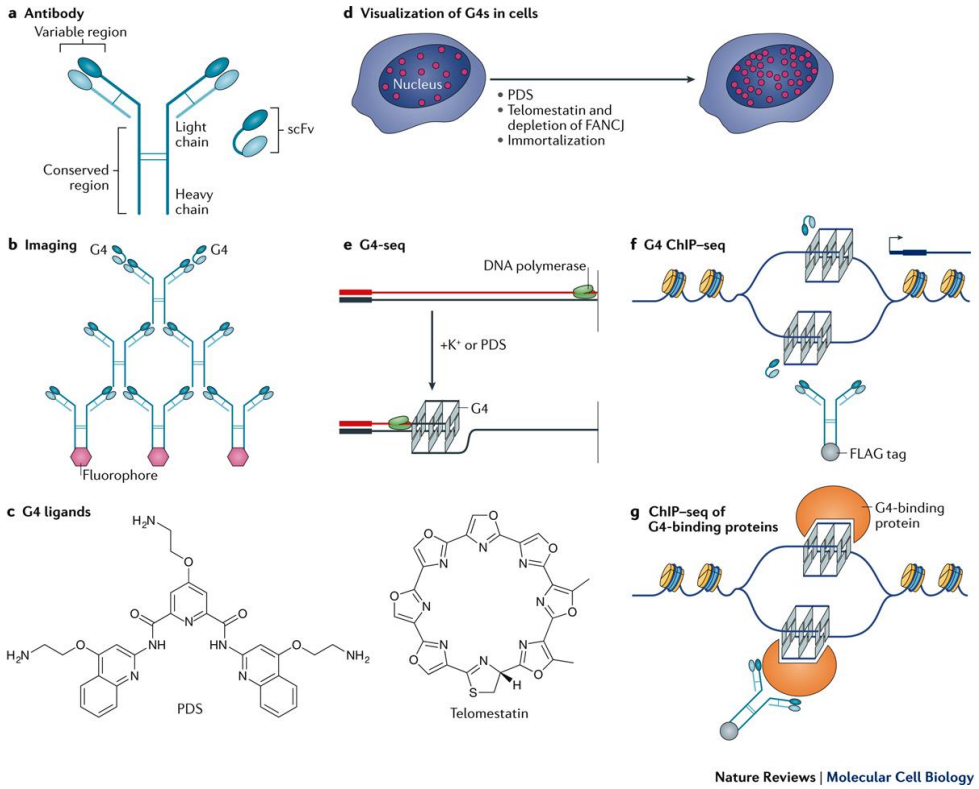

Figure 2 | G4 的主要侦测手段[4]。

从前往后依次是：抗体结构示意图；细胞内 G4s 的荧光定位；荧光标记 G4 的机理；G4-seq 技术图解；G4 ChIP-seq 技术图解；G4 配体的两个例子及其化学结构式；通过 G4 结合蛋白作用的 ChIP-seq。

3.1 G4 ChIP-seq[6-7]

对于细胞内特定结构或者特定序列的 DNA，目前已经掌握了一些蛋白质结合捕获的手段，也就是蛋白质免疫共沉淀技术。而相对应于 G4 结构设计特定的蛋白质作为探针，捕获序列后洗去蛋白质进行测序，我们称之为 G4 ChIP-seq。

G4 ChIP-seq 挑选 G4 选择性的探针去捕获 G4s，然后用抗体结合探针，通过蛋白免疫共沉淀与其他序列分离，然后去除蛋白质（探针），对序列进行高通量测序。通常，探针分为两类——抗体(antibodies)以及配体(ligands，存在部分配体并非蛋白质，而是小分子，但应用的原理基本一致)，这些探针具备蛋白质特异性识别的能力，能够识别特定的 G4 序列或者 G4 拓扑结构。

但是，也正是由于蛋白质的特异性，G4 ChIP-seq 往往针对的是典型的 G4 结构，测序结果也因选取的探针的不同而大同小异，这使得 G4 ChIP-seq 探测得到的结果并非完整的细胞内 G4 图谱。而且探针结合到特定的位点很有可能改变结构的稳定性，从而导致细胞内一些不可控的扰动，所以 G4 ChIP-seq 技术不太可能作用于微扰动实验情况下的细胞内 G4 位点的探究。同时，由于低复杂度序列区域的 GC 富集等问题，导致这些序列区域的 G4 ChIP-seq peak 存在一定的不稳定因素，需要在分析过程中剔除。

受限于生物种类以及细胞类别的多样性和复杂性，G4 ChIP-seq 目前还未能实现细胞异质性的数据累积，已经生成的数据有限，大部分都是针对人体细胞进行的测序。

装

订

线

3.2 G4-seq[6]

G4-seq，是一种结合了二代测序技术[8]和聚合酶停滞模型[9]的新型高通量测序技术。由于 G4 空间拓扑结构的特异性和稳定性，单链上的 G4 结构会导致聚合酶链式反应中的聚合酶停滞，进而影响碱基的召唤，导致聚合酶链式反应出错。研究人员比较了 G4 影响导致聚合酶停滞的序列和没有 G4 影响下的正常序列的测序质量和碱基含量，发现在 G4 位点（聚合酶停滞）的碱基错配概率极高，测序质量极低。于是，研究人员通过体外给定 G4 稳定条件来产生一条 G4 结构阻碍聚合酶的高错配序列，与正常条件下完成聚合酶链式反应的低错配序列比较，找到序列上特定的 G4 结构位点。

G4-seq 的一般流程如下：在 Na^+条件下对 DNA 序列进行准确的测序和比对，获得比对结果 read1，然后在 K^+ & PDS（促进 G4 稳定）条件下重新测序原模板，获得比对结果 read2，由于 G4 诱导聚合酶停滞改变了从 G4 结构开始的测序读数，导致 read2 的测序质量下降从而识别 G4 位点。事实上，G4-seq 正是采用了 mismatch 较多的序列来鉴定 G4s 位点，这和我们以往丢弃测序质量低以及错配高的序列的想法大相径庭，但是却极其有效。

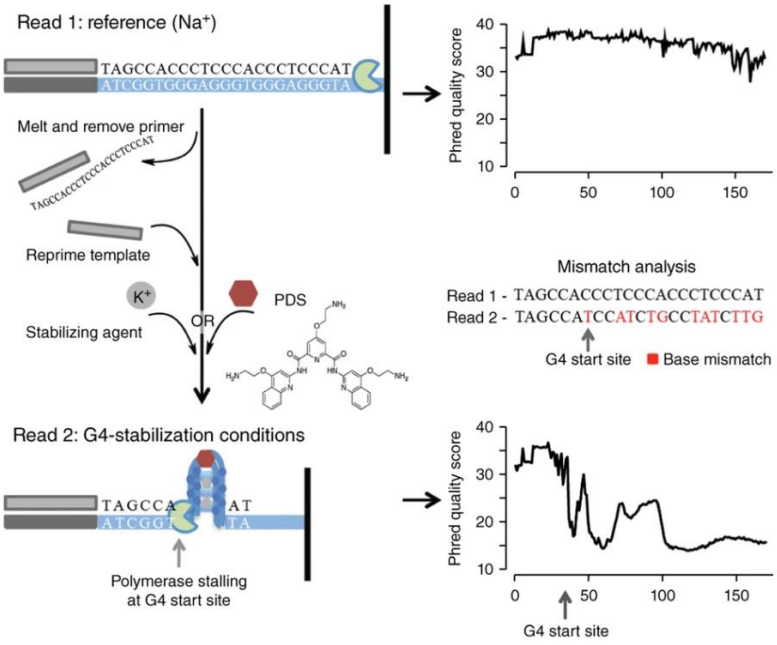

Figure 3 | G4-seq 的技术原理[6]。

Read 1 是在 Na+条件下对 DNA 序列进行准确测序比对得到的结果，用作为参考；Read 2 则是通过 K+或者 PDS 等条件稳定 G4 结构后对同样一条 DNA 序列进行测序比对的结果，由于 G4 结构稳定，聚合酶停滞，导致从 G4 起始位点开始会出现大量的碱基错配以及测序质量下降的情况。最终通过比较 Read 1 和 Read 2，我们就能找到 G4 结构位点。

G4-seq 关注于能够形成 G4 结构的位点，既避免了完全依靠序列不依赖结构稳定性的漏洞，又解决了 G4 ChIP-seq 特异性过高的问题，使得测序获得的 G4s 结构位点数目和 G4 种类都大大提升。同时，因为针对 G4s 结构与聚合酶停滞关系这一共性，G4-seq 能够很好的适用于多种生物的多种细胞的基因组。目前已经有实验给出了 12 个物种的 G4-seq 全基因组图谱数据[8]。G4-seq 的数据也往往由于准确性较高，被广泛用于计算预测方法的验证。

3.3 其他物理手段

一直以来，物理检测技术是作为结构探测的重要手段之一。这些实验技术手段能够用来检测特定序列的 G4 形成能力，而且通常都是可靠的。比较常见的技术[8]如核磁共振(NMR)，X-ray 晶体学等，它们可以用来检测结构信息；紫外线融化，用来探测 G4 的热稳定性；以及荧光标记等技术。但是，物理手段往往不具备高通量检测的能力，在鉴定全基因组层面的 G4 结构位点时并不如测序以及计算方法高效。

3.4 计算预测方法

除了传统意义上的实验探测手段，我们还能通过开发计算算法来预测 G4s 结构位点。通常来讲，计算方法假定一种 G4 的序列形式，这种形式是依据先验经验产生的，转化为计算语言后对基因组或者特定的数据集进行检索比对，最终获取计算机预测的 G4 序列及其位点。即使计算方法并没有直接采取实验手段，但是其理论的建立以及数据的来源都离不开实验，甚至于最终对结

装 订 线

果的验证以及对模型的完善都需要很好的实验数据支撑。

具体的算法我们将会在下文介绍。

4 计算机预测方法介绍

由于 G4 ChIP-seq、G4-seq 等实验技术的数据产生，研究 G4 的过程中积累了大量的先验经验和得到验证的数据集。研究者们希望能够借助计算机，大通量地获取全基因组的 G4s 结构位点。他们根据 G4s 的特定序列模式以及在序列上的分布特点，开发出了一些算法和工具。现有的主要算法和工具可以分为三大类[10]：通过 G4 序列模式进行正则匹配、滑动窗口结合打分矩阵进行评估预测以及机器学习的方式。

4.1 正则表达匹配

计算机的正则表达式指的是一段能够指代某一种模式的不连续字符串，经常用于检索。在计算生物学上，我们根据实验经验，归纳出 G4 序列具备的一些共性——由长度至少为 1 nt 的 loop 连接起来的四个鸟嘌呤 runs，也就是之前提到的典型的 G4s 基序：$G_{\geq 3}N_{1\text{-}7}G_{\geq 3}N_{1\text{-}7}G_{\geq 3}N_{1\text{-}7}G_{\geq 3}$。我们假定基因组上的 G4 都具备这一共性，设计出计算上的正则表达式，然后针对全基因组的序列进行检索，找出符合特征的序列以及位点。

这种检索的模式定义严格，输出结果单一，无法在序列和结构之间做出很好的平衡（过于严格的匹配模式找到的序列在体内不一定具备稳定的结构，但是过于宽泛的匹配模式又不能很好地剔除假阳性目标）。在应对基序或者 pattern 确定的序列结构的检索时，这种算法往往具有比较好的表现，但是在应对结构和序列复杂多变的 G4 时，这种算法并不是最优的选择。

4.2 滑动窗口与打分矩阵

为了能够应对 G4 序列的更多可能性，研究者们采用更宽泛的基序($G_{x1}N_{L1}G_{x2}N_{L2}G_{x3}N_{L3}G_{x4}$)进行检索，然后设定一定的规则，针对检索得到的序列进行打分和排序。设定的打分规则更多地考虑到 G4 结构的稳定性以及在体内形成的合理性。相比较于经典的正则表达匹配算法，打分的方式适当放宽了初步检索时对 G4 序列组成成分的要求，而且在后续打分过程中考虑到了结构稳定等要素，更加贴切于 G4s 结构的多层性。但是，这种基序的模式依然定义了严格的边界以及序列组成，不具备识别非典型 G4s 的能力。

于是研究者们将滑动窗口的策略结合到打分模式上。滑动窗口的策略主要关注三个参数：鸟嘌呤 run 的长度，滑动窗口的大小以及每次滑动的间距。在窗口大小为 w 时，检索器从序列起点出发，每次移动 s（间距）个核苷酸，检索包含 4 个长度至少为 k（runs 的长度）的 k-mer 鸟嘌呤的序列。最终，检索器就输出了很多包含四个 k 以上鸟嘌呤 runs 的窗口，这些鸟嘌呤 runs 至少被 1 个核苷酸分隔。滑动窗口的策略限制了 G4 候选序列的最大长度，而不是 G4 候选序列内部结构的特定长度，比较于之前的基序匹配模式，在泛用性上更胜一筹。其中比较具有代表性的算法工具有 G4Hunter[11-12]和 pqsfinder[13]等。

为了使得打分的规则更加严谨，通常我们都希望减少人为先验归纳的参数误差，采取实验数据验证和校准参数的方式完善模型。目前大部分的 G4s 结构位点计算预测工具都引入了体外验证

过的 G4-seq 数据集来检验模型的好坏，这使得在已知数据集下，工具的实用性得到了提升。

4.3 机器学习

上述两种计算方法大多基于经验以及专家知识，而且往往依靠人力总结有限数量的已观测到的 G4 结构的共性，无法完美地描绘 G4 多样性构象的完整图像。硬件技术的提升使得机器学习的方法逐渐热门，机器学习依靠数据本身驱动预测，避免了人为假设，往往更能够契合 G4 的序列特征。

机器学习方法[14]通常采集一批有效的实验数据，数据包含了一定量的阳性样本和阴性样本。通过设计回馈型模型，让计算模型从数据内部学习特征，然后使用这些特征去预测数据，最终使得在这批数据中能够得到最高的准确率。另外，为了避免机器学习模型过度依赖于一批数据，往往需要扩大学习的数据规模，或者增加一个测试数据集。当模型同时在训练数据集和测试数据集都表现良好时，我们认为这个学习器初步完成。但是，比较遗憾的是，机器学习学习出来的特征通常表现为计算语言的形式，而且目前没有将其解读为生物学特征的渠道，这使得机器学习获得的特征晦涩难懂，无法从生物学理论上进行解释，这也就是我们常说的黑箱(black-box)模型。

机器学习方法种类繁多，它通过训练一定量的学习器来学习样本的特征，最终实现回归或者分类，典型的算法有决策树、逻辑回归、支持向量机，前馈型神经网络（G4RNA screener[15]）等。在 python 语言中，已经有集成了这些算法的一个开源软件包 Scikit-learn[16]，能够直接使用。

理论上来讲，只要数据的质量（G4 种类）和数据的量（G4 位点的数量）足够，机器学习模型能够预测所有的 G4 结构位点，而且往往都很准确。这极大地避免了传统模式下的人为误差，降低了假阳性率和假发现率。但是，到目前为止，我们并不具备可靠的充足的 G4 结构位点数据，已有的已验证的数据集都比较小，而且没有办法扩展到非典型 G4 序列（没有足够的训练样本）上，这使得机器学习方法暂时还不太实用。

4.4 计算方法总结

前面我们提到了三种计算预测的手段，其中前两者基于实验经验，根据之前实验研究获取的 G4 的特定序列模式，特殊结构位点等信息来进行模型的构建；而后者则完全由数据驱动，基于大量的标记好的序列训练预测模型，达到最终的预测效果。但是，这些算法各自因为匹配模式的宽泛程度在不同细胞间存在差异以及不同细胞系的数据的数量和质量有所差异等种种原因，很难针对不同的细胞中的 G4s 结构位点进行准确预测。

然而，我们希望能够找到一种算法模型，可以解析不同组织细胞中 G4s 的结构位点图谱及其异质性，并且提升 G4s 结构位点预测的准确性。所以，发掘新的算法和数据，是研究 G4 位点预测的关键。

而且，随着研究的深入，科学家们已经不满足于探测 G4 初级序列（也就是单个 G4 结构），他们希望能够用计算方法直接预测更高阶的 G4 结构（由多个 G4 亚基组成的 G4 聚合结构），以便解析基因组或者 DNA 上更加复杂的结构和变化。目前这些算法尚在起步阶段，与预测非典型 G4 结构位点相类似，这部分算法缺乏足够的实验数据支撑以及实验验证，泛用性还不足以用来预测全基因组上的 G4 高阶结构。

5 总结和展望

我们整理了 G4 结构的特点，具体去了解了如何探测基因组上的 G4 结构以及能够辨别一个序列是否可以形成 G4 结构的方法。针对这些研究，我们提出了我们的想法，希望能够引进新的数据，进一步完善 G4 细胞异质性图谱的绘制，提高计算预测的准确性。

从聚合酶停滞模型我们可以看出，体外稳定的 G4 结构可以影响 DNA 复制和转录，导致 DNA 链的碱基错配，进一步可能引起损伤修复等机制（如果发生在体内）。不难想象，体内的 G4 结构也由于其稳定性以及在特定区域的富集程度，极有可能在表观遗传和 DNA 结构上具有一定的影响效应，包括 DNA 复制、转录、基因表达、基因组的稳定性以及表观遗传记忆等方面。

一些病理学研究表明，G4 和癌基因表达存在一定的关联，对于肿瘤学研究具有一定的价值。即便现在我们未能具体明了地解析这种关联，理解 G4 的机制，但是随着研究的进展，G4 一定能够展现出其临床价值和生物学意义。

装

订

线

参考文献

[1] Bochman ML, Paeschke K, Zakian VA. DNA secondary structures: stability and function of G-quadruplex structures [J]. Nat Rev Genet. 2012 Nov; 13(11):770-80.

[2] Rhodes D, Lipps HJ. G-quadruplexes and their regulatory roles in biology [J]. Nucleic Acids Res. 2015 Oct 15; 43(18):8627-37.

[3] Mukherjee AK, Sharma S, Chowdhury S. Non-duplex G-Quadruplex Structures Emerge as Mediators of Epigenetic Modifications [J]. Trends Genet. 2019 Feb; 35(2):129-144.

[4] Hänsel-Hertsch R, Di Antonio M, Balasubramanian S. DNA G-quadruplexes in the human genome: detection, functions and therapeutic potential [J]. Nat Rev Mol Cell Biol. 2017 May; 18(5):279-284.

[5] Kwok CK, Merrick CJ. G-Quadruplexes: Prediction, Characterization, and Biological Application [J]. Trends Biotechnol. 2017 Oct; 35(10):997-1013.

[6] Chambers VS, Marsico G, Boutell JM, et al. High-throughput sequencing of DNA G-quadruplex structures in the human genome [J]. Nat Biotechnol. 2015 Aug; 33(8):877-81.

[7] Hänsel-Hertsch R, Beraldi D, Lensing SV, et al. G-quadruplex structures mark human regulatory chromatin [J]. Nat Genet. 2016 Oct; 48(10):1267-72.

[8] Meyer CA, Liu XS. Identifying and mitigating bias in next-generation sequencing methods for chromatin biology. Nat Rev Genet. 2014 Nov;15(11):709-21.

[9] Han, H., Hurley, L.H., Salazar, M. A DNA polymerase stop assay for G-quadruplex-interactive compounds. Nucleic Acids Research, 27 (2), pp. 537-542, 1999.

[10] Puig Lombardi E, Londoño-Vallejo A. A guide to computational methods for G-quadruplex prediction [J]. Nucleic Acids Res. 2020 Jan 10; 48(1):1-15.

[11] Brázda V, Kolomazník J, Lýsek J, et al. G4Hunter web application: a web server for G-quadruplex

prediction [J]. Bioinformatics. 2019 Sep 15; 35(18):3493-3495.
[12] Brazda V, Kolomaznik J, Mergny JL, et al. G4Killer web application: a tool to design G-quadruplex mutations [J]. Bioinformatics. 2020 May 1; 36(10):3246-3247.
[13] Hon J, Martínek T, Zendulka J, et al. pqsfinder: an exhaustive and imperfection-tolerant search tool for potential quadruplex-forming sequences in R [J]. Bioinformatics. 2017 Nov 1; 33(21):3373-3379.
[14] 周志华. 机器学习. 北京: 清华大学出版社, 2016: 57-59
[15] Garant JM, Perreault JP, Scott MS. G4RNA screener web server: User focused interface for RNA G-quadruplex prediction [J]. Biochimie. 2018 Aug; 151:115-118.
[16] Pedregosa, F. and Varoquaux, G. and Gramfort, et al. Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.

装
订
线